Osteoarthritis and Cartilage (2008) **16,** 551–559

doi:10.1016/j.joca.2007.12.016

Osteoarthritis and Cartilage

International Cartilage Repair Society

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative[1]

A. M. Davis Ph.D., Dr.†‡§‖¶#*, A. V. Perruccio M.HSc.††††, M. Canizares M.Sc.††, A. Tennant Ph.D., Dr.‡‡, G. A. Hawker M.D., M.Sc., F.R.C.P.C., Dr.§§‖‖¶, P. G. Conaghan M.B.B.S., Ph.D., F.R.A.C.P., F.R.C.P., Dr.¶¶, E. M. Roos P.T., Ph.D., Dr.##†††, J. M. Jordan M.D., M.P.H., Dr.‡‡‡§§§, J.-F. Maillefert M.D., Ph.D., Dr.‖‖‖¶¶¶, M. Dougados M.D., Dr.### and L. S. Lohmander M.D., Ph.D., Dr.##

† *Division of Health Care and Outcomes Research, Toronto Western Research Institute, Toronto, Canada*
‡ *Arthritis Community Research and Evaluation Unit, Toronto Western Research Institute, Toronto, Canada*
§ *Department of Physical Therapy, University of Toronto, Toronto, Canada*
‖ *Department of Rehabilitation Science, University of Toronto, Toronto, Canada*
¶ *Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada*
# *Institute of Medical Science, University of Toronto, Toronto, Canada*
†† *Department of Public Health Sciences, University of Toronto, Toronto, Canada*
‡‡ *Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, Leeds, UK*
§§ *Division of Rheumatology, Department of Medicine, Women's College Hospital, Toronto, Canada*
‖‖ *Department of Medicine, University of Toronto, Toronto, Canada*
¶¶ *Academic Unit of Musculoskeletal Disease, University of Leeds, Leeds, UK*
Department of Orthopaedics, Clinical Sciences Lund, Lund University, Lund, Sweden
††† *Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark*
‡‡‡ *Department of Medicine, Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC, USA*
§§§ *Department of Orthopaedics, Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC, USA*
‖‖‖ *Dijon University Hospital, Dijon F-21079, France*
¶¶¶ *INSERM U887, University of Burgundy, Dijon F-21079, France*
Department of Rheumatology, Faculty of Medicine, Cochin Hospital, Paris, France

Summary

Objective: To derive a cross-culturally valid, short measure of physical function using function subscales (daily living and sports and recreation) of the Hip disability and Osteoarthritis Outcome Score (HOOS).

Methods: Rasch analysis was conducted on data from individuals from multiple countries who had hip osteoarthritis (OA). Fit of the data to the Rasch model was evaluated by model χ^2 and item fit statistics (χ^2, size of residual, and *F*-test). Differential item functioning was evaluated by gender, age and country. Unidimensionality was evaluated by factor analysis of residuals. Individual data sets were analyzed and data pooled and re-analyzed for fit to the model. Regression modeling was conducted to derive a nomogram converting raw summed scores to Rasch derived interval scores.

Results: Seven data sets were included ($n = 2991$), ages 19–96 years, male/female ratio was 1:1.23. The final model included five HOOS items. From the easiest to most difficult, the items (logit) were as follows: sitting (1.832), descending stairs (0.729), getting in/out of bath or shower (0.255), twisting/pivoting on loaded leg (−0.221) and running (−2.595). The separation index was 0.80.

[1]New Emerging Team Grant in Early OA sponsored by Canadian Arthritis Network and Canadian Institutes of Health Research, Pfizer, Novartis, Negma, Astra Zeneca, Rottapharm, Expansciences.AV Perruccio is funded by a CIHR Canada Graduate Scholarship.There was no involvement of the funding sources in the study design, data collection, analysis and interpretation of the data, in the writing of the manuscript or in the decision to publish the work.

*Address correspondence and reprint requests to: Aileen M. Davis, Ph.D., Senior Scientist, Division of Health Care and Outcomes Research, Toronto Western Research Institute, 399 Bathurst Street, MP11-322, Toronto, Ontario M5T 2S8, Canada. Tel: 1-416-603-5543; Fax: 1-416-603-6288; E-mail: adavis@uhnresearch.ca

Received 20 December 2007; revision accepted 26 December 2007.

Conclusion: The daily activity and sports and recreational items of the HOOS were reduced to five items achieving a feasible, short measure of physical function with interval level properties. This tool has potential for use as the function component of an OA severity scoring system. Further testing of this measure is warranted.

Key words: Osteoarthritis, Physical function, Outcome measure, Rasch analysis.

Introduction

Arthritis, particularly osteoarthritis (OA), ranks among the most prevalent diseases in the developed world and is a major cause of pain and physical disability[1,2]. It is most common in the hip and knee and is a leading cause of activity limitation, loss of independence, decreased quality of life and is a significant economic burden in terms of health care costs[1,3–6].

Total joint replacement (TJR) is a known effective treatment option for end stage hip or knee OA[7]. However, while studies have evaluated interventions for relieving pain and improving function, there has been little work in understanding interventions that might improve pain and functional disability in those who have mild or moderate symptomatic OA. Disease modifying agents for OA [Disease Modifying Osteoarthritis Drugs (DMOADs)] are of interest[8], but to evaluate these agents, there is a need to define eligibility criteria for clinical trials and appropriate outcomes.

As described by Gossec *et al.*[9], an international working group created under the auspices of Osteoarthritis Research Society International (OARSI) and Outcome Measures in Rheumatology Clinical Trials (OMERACT) determined that TJR, while the definitive outcome of failure in treatment of hip or knee OA, was not a feasible outcome in trials of non-surgical management, given the issues of access to TJR[10–12] and people's variability in willingness to undergo such surgery[13,14]. Hence, it was decided by the working group that the domains of pain, physical function and joint structure would be combined as a surrogate measure of outcome. Given this objective, it is critical that we have a parsimonious set of cross-cultural items that represent the range of difficulties of individuals across the spectrum of OA severity (that is, community dwelling individuals through to those with severe OA such that they are candidates for TJR). A working group was created for each domain, with the goal of determining a measure that would be integrated into the combined surrogate outcome. The focus of this paper is the physical function domain for hip OA.

The most common measures of physical function with demonstrated reliability and validity that have been used world-wide for hip OA include the Western Ontario McMaster Universities' Osteoarthritis Index (WOMAC)[15–17] and the Hip disability and Osteoarthritis Outcome Score (HOOS)[18]. The WOMAC physical function subscale includes 17 items that were selected based on their level of importance to people with hip and knee OA[15–17]. There is concern, however, that the WOMAC 3.0 physical function subscale does not include items of sufficient difficulty[19,20]. To address this limited range, the HOOS was developed[18]. The HOOS is inclusive of the 17 physical function items of the WOMAC but also includes higher demand function, sport and recreational activities, increasing the physical function items to 21.

Additionally, concerns have been expressed that the WOMAC physical function subscale has redundancy within its restricted range of difficulty given the number of items and the method of determining their inclusion[19,21]. Reduced item sets for the WOMAC physical function subscale have been developed and tested by Whitehouse *et al.*[22] and by Tubach *et al.*[23]. Whitehouse *et al.*[22] reduced the 17-item physical function subscale to seven items based on the opinion of 36 orthopedic and rheumatology personnel from the United Kingdom and United States. These opinion leaders were asked to indicate the five items they would keep based on items most likely to change after joint replacement surgery, what patients cared about most, and items that represented a broad spectrum of difficulty. The measurement properties subsequently were tested in data from patients from the United Kingdom, United States and Australia who had total knee replacement. Tubach *et al.*[23] asked 1362 patients with hip or knee OA to select the five items representing activities that they felt needed to improve the most and 399 rheumatologists were asked to select five items that they considered resulted in the most problems for people with hip or knee OA. Based on their analyses eight items were chosen that were considered most important to the patients. These authors similarly tested the measurement properties of the shortened scale. The difficulty with these approaches is that neither included the spectrum of severity of OA in their work, the WOMAC items with a limited range of difficulty were used, and the scaling properties of the items were not considered in the process.

Item response theory methods, specifically the Rasch model, have been used to develop and internally validate short measures[24]. The Rasch model[25,26] uses a logistic function that creates an interval-scaled measure. The standard error (SE) of an item is independent of the SE of other items such that there should be improved accuracy and stability of the performance of the items across different populations. Additionally, the item difficulty parameters explicitly demonstrate the range of difficulties (in this case the range of functional difficulty) that are represented by the measure.

We, therefore, secondarily analyzed HOOS and WOMAC 3.0 data from individuals from Europe and North America with hip OA accrued to community cohorts or individuals about to have total hip replacement surgery using the Rasch model to develop a short measure of physical function.

Methods

Data ($n = 2991$) from community (Canada) and pre-total hip replacement surgery cohorts (four from Canada, one from Sweden and the Eurohip data set which represents data from Austria, Finland, France, Germany, Hungary, Iceland, Italy, Poland, Spain, Sweden, Switzerland and the United Kingdom) were included in the analysis (Table I). The methods for accrual to the community sample have been described elsewhere[10,27]. For those pre-THR, all patients were booked for their surgery and completed the questionnaires either as a part of routine care or in relation to a specific research study. The age of the total sample ranged from 19 to 96 years with four data sets including individuals with an average age in the 60s. There were slightly more females than males. All data analyzed were based on the WOMAC Likert-type version 3.0 or HOOS Likert-type version 2.0 questionnaires. This secondary analysis was approved by the institutional ethics review board.

MEASURES

The WOMAC version 3.0[15,17] physical function subscale consists of 17 items scored 0–4 with response options for rating the amount of difficulty on an activity ranging from 'None' to 'Extreme'. The subscale score is calculated by summing the raw responses with a score range of 0–68 where high

Table I
Description of the data sets used in the analyses

Data set	Country	Type of sample	*N*	Age, mean (SD, range)	Sex*, M:F
1	Canada	Community	122	77.1 (7.0, 65–96)	82:40
2	Canada	Pre-THR	371	69.9 (12.0, 31–92)	122:143
3	Canada	Pre-THR	1078	63.1 (14.0, 19–96)	462:616
4	Canada	Pre-THR	78	61.5 (13.3, 28–86)	30:46
5	Canada	Pre-THR	205	70.6 (10.8, 43–92)	44:123
6	Sweden	Pre-THR	92	71.0 (8.1, 51–88)	43:46
7	Eurohip	Pre-THR	1045	67.8 (10.8, 28–94)	491:554

*M:F ratio does not equal the sample size in some instances due to missing data.

scores indicate more difficulty. This WOMAC 3.0 subscale is included in the HOOS as the Function, daily activities scale[18]. Four additional Function, sports and recreational items of the HOOS[18] are similarly rated and scored to provide a raw subscale score of 0–16. For the HOOS subscales, the raw subscale scores are then calculated as a percentage score. The raw responses of the 21 items for the two HOOS subscales were used for these analyses.

ANALYSIS

The most basic form of the Rasch model, based on a dichotomous response scale, is that the probability of a person endorsing an item is a logistic function of the difference between the person's ability and the difficulty of the item. This can be expressed as a logit model as follows:

$$\ln\frac{p_{ni}}{1-p_{ni}} = \varphi_n - b_i$$

where ln is the logarithm function, P is the probability of person n endorsing item i, φ_n is the level of functional ability of person n, and b_i is the difficulty of item i. The item and person estimates are expressed as logits which allow for linear transformation of the raw score. As an item estimate is based on responses to the other items, the model is able to accommodate missing responses to an item for a given respondent. An extension of this model, the partial credit model for multiple-response option data, the details of which are described elsewhere[28], was applied in this analysis using RUMM2020 software[29]. This form of the logistic model addresses the case of multiple-response option data such as that of the HOOS where the response represents the individual's difficulty rather than a "correct" answer.

The criteria for interval level data include: demonstration of appropriate response category ordering, fit of the data to the model, lack of item bias or differential item functioning (DIF) and unidimensionality[25,30–33]. The data were considered to fit the Rasch model when the item χ^2 probability was not significant, the item residuals were small (i.e., absolute value smaller than 2.5) and the F-test statistics were not significant. Statistical significance was based on a critical value of 0.05 with a Bonferroni correction factor for multiple testing[34] with P-values < 0.002 considered statistically significant.

Response categories were examined to determine if they produced sequentially ordered item thresholds. Thresholds are the point between two response categories where there is 50% probability of either response. Where items had misordered thresholds, the response categories were collapsed.

Item bias or DIF occurs when there are systematic differences in responses based on characteristics of the respondents. Invariance of the model by age, sex, and country was evaluated. Although item splitting can be done to address DIF by having different difficulty estimates or logit values for an item, we *a priori* determined that items with DIF would be removed to maximize parsimony.

The internal consistency and reliability of the final model are evaluated by a separation index that is equivalent to Cronbach's alpha[35]. Values of approximately 0.80 and greater are acceptable[36].

The assumption underlying the Rasch model is that the items form a unidimensional scale. Most commonly, unidimensionality is assessed by performing principal component analysis of the residuals. There should be no factor structure to the residuals as the Rasch analysis has 'extracted' the factor for which item associations exist. Interpretation of these results can be difficult as the residuals represent an unknown amount of the total variance as presented by the RUMM software. Care must be taken to avoid interpretation of a suggested data pattern based on the residuals of one or two items that provides minimal explanation of the total variance[33]. As a final test of unidimensionality, person score estimates were compared based on subsets of items from the factor analysis of the residuals. Scores were created from the items with positive factor loadings of 0.30 and higher and also from items with negative factor loadings of −0.30 and lower. T-tests were then done to compare the estimates for each person and the percentage of tests outside ±1.96 (95% confidence interval) was calculated[31,33].

Person score estimates are provided on a logit scale. We *a priori* expected that the distribution of the community sample would demonstrate a proportion of individuals with less disability than the samples of individuals who were waiting for THR.

Finally, we used regression methods to determine how well the five-item interval level scale predicted the scores the raw summed score of the five items by fitting a cubic model. To facilitate ease of use in clinical and research settings, we then created a nomogram equating the raw sum score to the interval level measure scaled from 0 to 100.

Results

Table II presents an overview of the model summary statistics for some of the Rasch analyses. The first analytic model was run with all 21 items using only the Canadian data sets from both the community and the pre-total hip replacement samples. The mean item location for the initial model based on 21 items was 0.00 with Standard Deviation (SD) = 1.341. The SD should be approximately 1 although values of approximately 1.4 are common. The overall model χ^2 statistic was significant with a P-value of <0.0001. Ten of the 21 items were problematic based on misfit criteria (Bonferonni corrected significant χ^2, large residuals, significant F-test) and are bolded in Table III. As an example, the item *walking on a flat surface* has a large residual of −6.670 which is well beyond the threshold magnitude of ±2.5 and the P-values are beyond the Bonferroni corrected critical value for both the χ^2 statistic and the F-test.

Given these findings, the pattern of the items thresholds was evaluated to see if there was disordering of responses for any items. The items *heavy domestic duties*, *running*

Table II
Summary of models fitted

Model	Sample size	# Items	Items				Persons				Item-trait inter-action		PSI
			Location		Fit residual		Location		Fit residual		χ^2	χ^2 prob	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD			
Initial	1636	21	0	1.341	−0.307	5.638	0.060	1.597	−0.409	1.614	1192.91	<0.0001	0.95
4	1613	11	0	1.736	0.057	1.078	0.245	1.715	−0.465	1.276	70.02	0.3442	0.92
10	2476	10	0	1.574	−0.188	2.021	0.736	1.627	−0.478	1.264	102.19	0.0006	0.90
18	2688	7	0	2.176	0.356	1.480	1.017	1.723	−0.499	1.134	81.59	0.0002	0.85
Final	2643	5	0	1.640	0.170	0.886	0.627	1.617	−0.513	0.990	42.29	0.0672	0.80

Table III
First model using only Canadian data with 21 items

Item	Location	SE	Residuals	χ^2	*P*-value	*F*-stat	*P*-value
Sitting	1.729	0.03	−2.063	13.1910	0.0401	2.567	0.0178
Rising from sitting	1.009	0.03	−0.683	12.3320	0.0550	2.262	0.0353
Walking on a flat surface	0.973	0.03	**−6.670**	52.2410	**0.0000**	12.88	**0.0000**
Bending to the floor	0.698	0.03	0.336	12.9720	0.0435	2.091	0.0515
Descending stairs	0.677	0.03	−2.213	24.1440	**0.0005**	4.578	**0.0001**
Getting in/out of car	0.670	0.03	−1.938	18.8970	0.0043	3.523	**0.0018**
Taking off socks/stockings	0.569	0.03	**−4.659**	23.2830	**0.0007**	5.608	**0.0000**
Getting on/off toilet	0.514	0.03	−2.409	10.7070	0.0979	2.154	0.0448
Standing	0.485	0.03	**−2.625**	15.6730	0.0156	3.4	0.0024
Going shopping	0.388	0.03	**−4.744**	31.9120	**0.0000**	6.695	**0.0000**
Putting on socks/stockings	0.387	0.03	**−7.844**	57.5280	**0.0000**	15.093	**0.0000**
Light domestic duties	0.327	0.03	**15.737**	466.5960	**0.0000**	52.079	**0.0000**
Lying in bed	0.309	0.03	**3.828**	44.4820	**0.0000**	6.37	**0.0000**
Heavy domestic duties	0.208	0.06	**13.803**	364.4350	**0.0000**	40.997	**0.0000**
Rising from bed	0.140	0.03	−1.608	7.3680	0.2881	1.621	0.1374
Ascending stairs	0.130	0.03	−1.073	10.6390	0.1002	1.875	0.0817
Getting in/out of bath/shower	0.082	0.03	0.833	5.1250	0.5279	0.932	0.4707
Twisting/pivoting on loaded leg	−0.745	0.11	0.111	3.6570	0.7230	0.664	0.6789
Walking on uneven surface	−1.413	0.11	−1.948	9.3010	0.1574	1.961	0.0737
Squatting	−3.271	0.29	−0.225	1.3490	0.9689	0.264	0.9525
Running	−3.865	0.30	−0.395	7.0750	0.3139	1.825	0.0984

Mean = 0.00; SD = 1.34; separation index = 0.95; $\chi^2 = 1192.91$; and *P*-value <0.0001.

and *squatting* had disordered thresholds and categories were collapsed to achieve sequential order. The remaining items had properly ordered thresholds and all response categories were maintained. Additionally, the threshold distances varied across items such that the partial credit model was appropriate for the analysis of these data (data not shown).

DIF based on age and sex was also evaluated to determine if this was contributing to the misfit of items. No attempt was made to improve fit to the model by item splitting (i.e., allowing different estimates for items based on the DIF characteristics examined) as noted in Methods. Misfitting items were systematically removed, resulting in 11 items that fit the model (Table IV). The overall model had an item mean of 0.00, SD = 1.74, with a χ^2 statistic with a *P*-value of 0.3442. These 11 items provided the basis as we continued to work toward a final model using the additional data sets. The additional data sets were added one at a time allowing us to systematically test for DIF by country.

Addition of the remaining data sets resulted in a final five items that fit the Rasch model (Tables II and V). The overall model had an item mean of 0.0, SD = 1.64, which is improved from the 11-item model. Given the multiple testing and multiple comparisons, the χ^2 statistics, residuals and *P*-values met the criteria for item fit. The exception was *descending stairs* which had a significant *F*-test. This suggests that the class intervals (i.e., the distance between the thresholds) vary in width. None of the five items demonstrated DIF by age, sex, or country based on a Bonferroni-adjusted *P*-value. As an example, Fig. 1a and b shows that the items *getting in and out of the bath/shower* and *sitting* have no DIF by age or sex as the curves overlap.

The five items included in the final model include three from the original WOMAC 3.0 physical function subscale (*sitting*, *descending stairs* and *getting in and out of the bath/shower*) and two items added in the HOOS (*twisting/pivoting on loaded leg* and *running*). The items have average logits ranging from 1.832 (*sitting*, which is the easiest item) to −2.595 (*running*, which is the most difficult item) representing a range of difficulty (Table V). The item thresholds (Fig. 2) also demonstrate the range of difficulty with thresholds ranging from −10 to 5. The separation index is 0.80.

Table IV
Final model with 11 items based on Canadian data

Item	Location	SE	Residuals	χ^2	*P*-value	*F*-stat	*P*-value
Sitting	2.019	0.04	−1.013	6.4780	0.3718	1.073	0.3762
Getting in/out of car	1.223	0.03	0.231	3.3400	0.7651	0.501	0.8081
Rising from sitting	0.872	0.04	−1.423	9.4620	0.1492	1.896	0.0782
Descending stairs	0.872	0.03	1.675	5.5090	0.4803	1.214	0.2961
Getting on/off toilet	0.709	0.03	0.733	12.8570	0.0454	2.453	0.0230
Standing	0.682	0.04	−0.283	6.3470	0.3855	1.382	0.2181
Ascending stairs	0.286	0.03	−1.030	13.7240	0.0329	2.742	0.0118
Getting in/out of bath/shower	0.247	0.03	1.879	4.6180	0.5936	0.806	0.5653
Twisting/pivoting on loaded leg	−0.374	0.11	0.508	2.9760	0.8119	0.516	0.7957
Squatting	−2.862	0.29	−0.137	0.9840	0.9862	0.171	0.9839
Running	−3.674	0.32	−0.511	3.7260	0.7137	0.902	0.4956

Mean = 0.00; SD = 1.74; separation index = 0.92; $\chi^2 = 70.02$; and *P*-value = 0.3442. *T*-test for unidimensionality: Proportion of *t*-tests $p < 0.05 = 7.20\%$.

Table V
All data combined for final model with five items

Item	Location	SE	Residuals	χ^2	*P*-value	*F*-stat	*P*-value
Sitting	1.832	0.026	1.370	10.12	0.1197	2.22	0.0388
Descending stairs	0.729	0.026	−0.526	15.09	0.0196	3.93	**0.0006**
Getting in/out of bath/shower	0.255	0.026	−0.695	9.73	0.1363	3.00	0.0063
Twisting/pivoting on loaded leg	−0.221	0.102	0.800	5.15	0.5253	1.15	0.3367
Running	−2.595	0.301	−0.099	2.20	0.9002	0.49	0.8121

Mean = 0.00; SD = 1.64; separation index = 0.80; $\chi^2 = 42.29$; and *P*-value = 0.0672. *T*-test for unidimensionality: Proportion of *t*-tests $p < 0.05 = 2.60\%$.

In strict testing of unidimensionality, factor analysis of the residuals demonstrated no pattern (data not shown). Additionally, neither subset of items from the factor analysis of the residuals demonstrated a significant difference from the person estimates from the full five-item measure based on a Bonferroni-adjusted *P*-value. Only 2.6% of the sample demonstrated differences, supporting a unidimensional construct (Table V).

Person score distributions are shown in Fig. 2 and range from −5 to 7 logits with the community sample, as expected, including individuals with less disability (lower scores represent less disability).

Finally, given that we significantly shortened the two subscales of the HOOS, we wanted to evaluate how the Rasch scale predicted the summed score of the five items. Figure 3 shows the relationship of the Rasch-based scores in relation to the summed score. Additionally, Fig. 4 shows the prediction line based on an estimation of the cubic function regressing the Rasch score on the sum of the five items in the final Rasch-based scale. Table VI shows the model summary and coefficient estimates as well as the descriptive statistics for the raw scores, observed and predicted Rasch-based scores for the five items. Figure 4 supports the data in Table VI and the appropriateness of the cubic function.

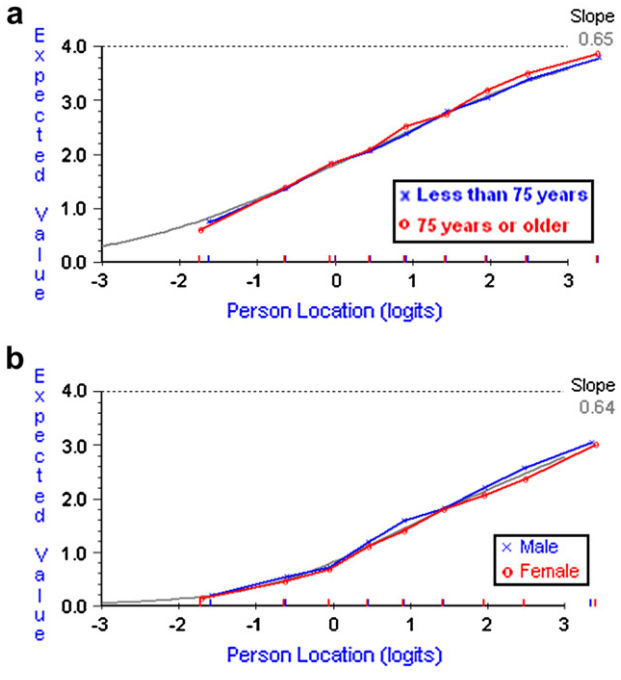

Fig. 1. Graphs demonstrating the lack of DIF for *getting in and out of the bath/shower* (a) for age and for *sitting* (b) by sex. For each graph, the *x*-axis shows the person score in logits and the *y*-axis represents the expected scores based on the Rasch model (overall fit is based on a χ^2 distribution). For each of *getting in and out of the bath/shower and sitting* the lines for age and sex, respectively, overlap demonstrating that there is no DIF by these factors.

By solving the cubic model in Table VI, Rasch-based, interval level scores can be estimated based on the summed score of the five items. Table VII presents these estimated Rasch-based scores for all possible integer values of the raw summed scores. The estimates are shown in both the original scale and rescored on a 0−100 scale where 0 represents no physical difficulty. As an example, a raw summed score of 13 is equivalent to a score of 50.8 on the interval score ranging from 0 to 100.

Discussion

Use of the Rasch model in this study supports a short measure, the five-item HOOS-PS (Appendix 1), for measuring physical function in people with OA of the hip. The measure covers a range of difficulty and represents a unidimensional construct and demonstrates no DIF by age, sex, or country despite the diversity in the test samples. As such a single score from the HOOS-PS can be created for an individual using Rasch methods and this score has interval level properties. This is supported by the evaluation of strict unidimensionality, which remains a challenge, but the methods used in this study represent current best methods[31,33].

One item, descending stairs, fit the model with the exception of the *F*-test criterion which suggests that there are statistically significant differences in the threshold distances for this item. The impact of one item failing a single criterion given that all other criteria for item and model fit were met, including the tests of invariance and strict unidimensionality, is unclear and currently under debate.

The other potential issue is that the Person-Separation Index (PSI) is 0.80 in the final model which is on the low end of acceptable levels particularly for individual level data[36]. Although the Cronbach's alpha (PSI in this case) is related to the intraclass correlation test−retest reliability coefficient[37], formal test−retest reliability with calculation of the appropriate test of concordance should be conducted for use of this measure in evaluating change.

The items derived from the WOMAC physical function subscale that are included in the HOOS-PS differ from those included in the short versions of the WOMAC developed by Whitehouse *et al.*[22] and by Tubach *et al.*[23,38]. *Sitting* is the only common item with the seven-item short WOMAC derived by Whitehouse *et al.*[22]. Their short measure includes the following WOMAC items: *ascending stairs*, *rising from sitting*, *walking on flat*, *getting in/out of car*, *putting on socks/rising from bed*, and *sitting*. Tubach

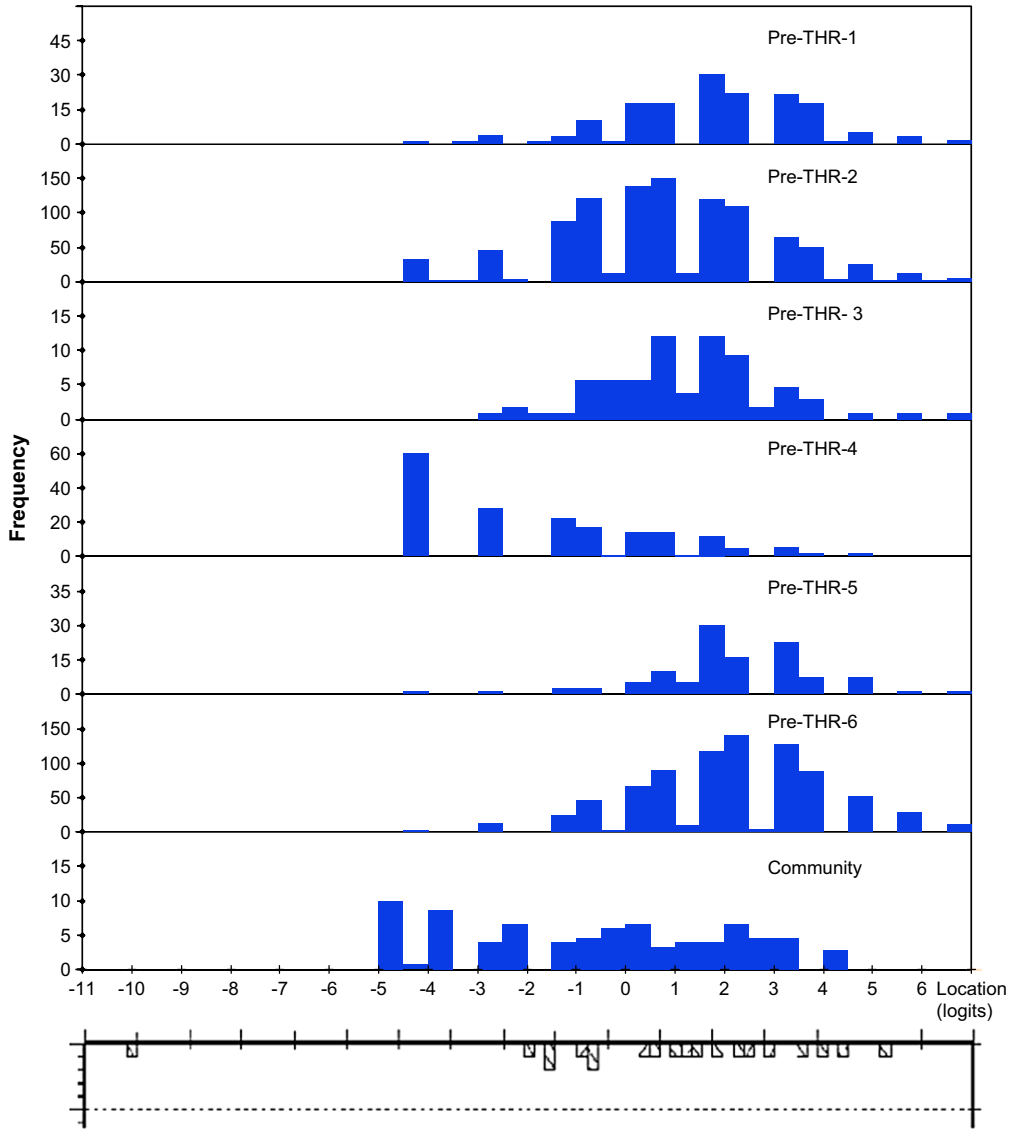

Fig. 2. Targeting map displaying both the range of item difficulties (hatched bars on the bottom) and the distribution of person abilities by sample (solid bars) over range of the logit values. The item logits range from −10 to 5 and the person scores range from −5 to 7. For each sample, the height of the solid bar represents the number of people who achieved a given person score.

et al. retained eight items of the WOMAC: *descending stairs*, *ascending stairs*, *rising from sitting*, *walking on flat*, *getting in/out of a car*, *going shopping*, and *putting on socks/stockings*[23]. Only *descending stairs* is common with the HOOS-PS. Given the different methods and different samples used by Whitehouse *et al.*[22] and Tubach *et al.*[23], as compared to the current work, these differences are not surprising. A major advantage of the current approach is that it achieves interval level measurement across a spectrum of OA disease severity and is free from DIF. Additionally, the HOOS-PS includes a greater range of item difficulty by its very inclusion of more demanding activities than those included in the WOMAC.

Short, as compared to long, versions of questionnaires that can be used for measuring patient status or change in status are in constant demand by clinicians, researchers and regulators to improve feasibility and compliance, especially when multiple questionnaires are being used. However, the major criticism of short questionnaires relates to their content validity. Content validity is a qualitative assessment determined mainly by how the items were generated. Ideally, the literature and all stakeholders are canvassed to determine that there are no critical omissions and that there is no irrelevant content[39] at the item generation phase of questionnaire development. In item reduction, content validity must be balanced with item redundancy and the additional information gained by each item.

The short versions developed by Whitehouse *et al.*[22] and Tubach *et al.*[23,38] have measurement properties of reliability, validity and responsiveness even though the items in the two short versions are very different. This suggests that the original 17-item scale had redundant items.

While aspects of truth, discrimination and feasibility as defined by the OMERACT filter[40] have been addressed in the current work further testing of the HOOS-PS is required. Generalizability of the current work is limited by

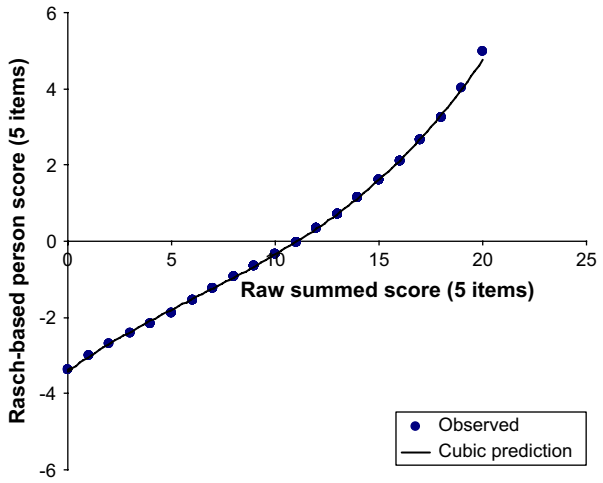

Fig. 3. The five-item raw summed scores (*x*-axis) are graphed against the Rasch-based person score in logits. The solid line shows the fit of a line to the data based on a cubic function.

the types of sample and cultural considerations of the countries from which data were available. Future studies will need to further evaluate reliability and validity and address cross-cultural validation. Additionally, given the intended use of the HOOS-PS as part of the composite index that will as one of its goals define an endpoint for those who have failed DMOAD therapy such that they are candidates for total hip replacement, validation studies to define a HOOS-PS cut-point will be required. Further applications of the HOOS-PS will require studies to define the responsiveness of the measure in different contexts in varying samples of people with hip OA who receive different treatment interventions.

In summary, based on accepted methods of measurement using Rasch analysis and using data from samples representing a spectrum of OA severity, we have developed a short measure of physical function, the HOOS-PS. This short measure fits the unidimensional, interval level scaled Rasch model and is free of DIF. Further, we have provided a conversion for raw summed scores to an interval scale for ease of use and interpretation in clinical and research settings.

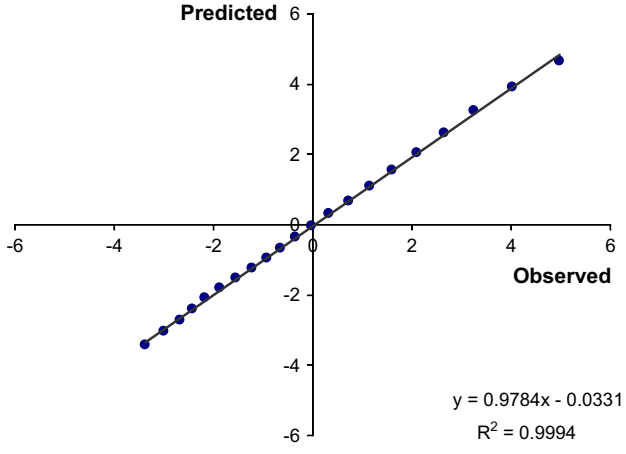

Fig. 4. Relationship between observed Rasch-based scores for final five items and predicted person scores obtained from cubic equation in Table VI.

Table VI
Raw-based scores (logits) regressed by raw summed score for final five items

	Model estimates			
	Coefficient	SE	*t*	*P*-value
Constant	−3.4104	0.0082	−416.54	<0.0001
(Raw summed score)^3	0.0009	0.0001	75.24	<0.0001
(Raw summed score)^2	−0.0171	0.0004	−47.24	<0.0001
Raw summed score	0.3851	0.0031	122.46	<0.0001
	Model fit			
	R	R^2	Adjusted R^2	
	0.999	0.999	0.999	
	Range			
	Minimum	Maximum	Mean	SD
Raw summed score	0	20	11.541	4.326
Observed Rasch-based person scores	−3.386	4.968	0.379	1.595
Predicted person scores	−3.410	4.549	0.289	1.552
Residuals	−0.062	0.418	0.000	0.054

Table VII
Crosswalk table of raw scores (0–20 to continuum scores, logits and 0–100 scale)

Total raw score (0–20)	Person interval level score in logits	Person interval level score (0–100 scale)
0	−3.4104	0.0
1	−3.0415	4.6
2	−2.7014	8.8
3	−2.3847	12.7
4	−2.0860	16.4
5	−1.7999	20.0
6	−1.5210	23.4
7	−1.2439	26.9
8	−0.9632	30.4
9	−0.6735	33.9
10	−0.3694	37.7
11	−0.0455	41.7
12	0.3036	46.1
13	0.6833	50.8
14	1.0990	55.9
15	1.5561	61.6
16	2.0600	67.9
17	2.6161	74.8
18	3.2298	82.4
19	3.9065	90.8
20	4.6516	100.0

Conflict of interest

None of the authors has any conflict of interest or disclosures to report in relation to this work.

Acknowledgments

The following individuals contributed data for this project: Aileen Davis (Canada), Paul Dieppe (Eurohip), Gillian Hawker (Canada), Stefan Lohmander (Sweden), Nizar Mahomed (Canada), Anna Nilsdotter (Sweden), Maria Suarez-Almazor (Canada), and James P. Waddell (Canada).

Appendix 1. HOOS-Physical Function Shortform (HOOS-PS)

This survey asks for your view about your hip. This information will help us keep track of how well you are able to perform different activities. Answer every question by ticking the appropriate box, only *one* box for each question. If you are unsure about how to answer a question, please give the best answer you can making sure you answer all the questions.

The following questions concern your level of function in performing usual daily activities and higher level activities. For each of the following activities, please indicate the degree of difficulty you have experienced in the *last week* due to your hip problem.

1. Descending stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

2. Getting in/out of bath or shower

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

3. Sitting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

4. Running

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

5. Twisting/pivoting on your loaded leg

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

References

1. Badley EM, Wang PP. Arthritis and the aging population: projections of arthritis prevalence in Canada 1991 to 2031. J Rheumatol 1998;25(1): 138–44.
2. Kaplan W, Laing R. Priority Medicines for Europe and the World. Geneva: WHO, Department of Essential Drugs and Medicines Policy; 2004.
3. Badley EM, Rasooly I, Webster GK. Relative importance of musculoskeletal disorders as a cause of chronic health problems, disability, and health care utilization: findings from the 1990 Ontario Health Survey. J Rheumatol 1994;21(3):505–14.
4. Raina P, Dukeshire S, Lindsay J, Chambers LW. Chronic conditions and disabilities among seniors: an analysis of population-based health and activity limitation surveys. Ann Epidemiol 1998;8(6):402–9.
5. Verbrugge LM, Patrick DL. Seven chronic conditions: their impact on US adults' activity levels and use of medical services. Am J Public Health 1995;85(2):173–82.
6. Perruccio AV, Power JD, Badley EM. The relative impact of thirteen chronic conditions across three different outcomes. J Epidemiol Commun Health 2007;61(12):1056–61.
7. Zhang W, Doherty M, Arden N, Bannwarth B, Bijlsma J, Gunther K-P, *et al*. EULAR evidence based recommendations for the management of hip osteoarthritis: report of a task force of the EULAR Standing Committee for International Clinical Studies Including Therapeutics (ESCISIT). Ann Rheum Dis 2005;64(5):669–81.
8. Buckwalter JA, Saltzman C, Brown T. The impact of osteoarthritis: implications for research. Clin Orthop Relat Res 2004;(427 Suppl): S6–S15.
9. Gossec L, Hawker G, Davis AM, Maillefert JF, Lohmander LS, Altman R, *et al*. OMERACT/OARSI initiative to define states of severity and indication for joint replacement in hip and knee osteoarthritis. J Rheumatol 2007;34(6):1432–5.
10. Hawker GA, Wright JG, Coyte PC, Williams JI, Harvey B, Glazier R, *et al*. Determining the need for hip and knee arthroplasty: the role of clinical severity and patients' preferences. Med Care 2001;39(3): 206–16.
11. Woolhead GM, Donovan JL, Chard JA, Dieppe PA. Who should have priority for a knee joint replacement? Rheumatology (Oxford) 2002; 41(4):390–4.
12. Maillefert JF, Hawker GA, Gossec L, Mahomed NN, Lohmander LS, Dieppe PA, *et al*. Concomitant therapy: an outcome variable for musculoskeletal disorders? Part 2: Total joint replacement in osteoarthritis trials. J Rheumatol 2005;32(12):2449–51.
13. Hudak PL, Clark JP, Hawker GA, Coyte PC, Mahomed NN, Kreder HJ, *et al*. "You're perfect for the procedure! Why don't you want it?" Elderly arthritis patients' unwillingness to consider total joint arthroplasty surgery: a qualitative study. Med Decis Making 2002; 22(3):272–8.
14. Ballantyne PJ, Gignac MA, Hawker GA. A patient-centered perspective on surgery avoidance for hip or knee arthritis: lessons for the future. Arthritis Rheum 2007;57(1):27–34.
15. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15(12):1833–40.
16. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically-important-patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1988;1:95–108.
17. Bellamy N. WOMAC Osteoarthritis Index A User's Guide IV. London, 2000.
18. Nilsdotter AK, Lohmander LS, Klässbo M, Roos EM. Hip disability and osteoarthritis outcome score (HOOS)—validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003;4:10.
19. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res 1999;12(5): 331–5.
20. Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, *et al*. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol 2003;56(11):1076–83.
21. Sun Y, Sturmer T, Gunther KP, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee—a review of the literature. Clin Rheumatol 1997;16(2):185–98.
22. Whitehouse SL, Lingard EA, Katz JN, Learmonth ID. Development and testing of a reduced WOMAC function scale. J Bone Joint Surg Br 2003;85(5):706–11.
23. Tubach F, Baron G, Falissard B, Logeart I, Dougados M, Bellamy N, *et al*. Using patients' and rheumatologists' opinions to specify a short form of the WOMAC function subscale. Ann Rheum Dis 2005;64: 75–9.
24. Cole JC, Rabin AS, Smith TL, Kaufman AS. Development and validation of a Rasch-derived CES-D short form. Psychol Assess 2004;16(4): 360–72.
25. Rasch G. Probabilistic Model for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press; 1960. (Reprint ed.).
26. Andiel C. Rasch analysis: a description of the model and related issues. Can J Rehabil 1995;9(1):17–25.
27. Hawker GA, Wright JG, Coyte PC, Williams JI, Harvey B, Glazier R, *et al*. Differences between men and women in the rate of use of hip and knee arthroplasty. N Engl J Med 2000;342(14):1016–22.
28. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47(2):149–74.
29. Andrich D, Lyne A, Sheridan B, Luo G. RUMM 2020. Perth: RUMM Laboratory; 2003.
30. Holland PW, Wainer H. Differential Item Functioning. Hilldale, NJ: Lawrence Erlbaum Associates; 1993..
31. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and component analysis of residuals. J Appl Meas 2002;3(2):205–31.
32. Pallant JF, Miller RL, Tennant A. Evaluation of the Edinburgh post natal depression scale using Rasch analysis. BMC Psychiatry 2006;6:28.
33. Tennant A, Pallant JF. Unidimensionality matters!. Rasch Meas Trans 2006;20(1):1048–51.
34. Bhakta B, Tennant A, Horton M, Lawton G, Andrich D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. BMC Med Educ 2005;5(1):9.
35. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297–334.
36. Nunnally JC. Psychometric Theory. New York: McGraw-Hill; 1978.

37. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol 1991; 44(4–5):381–90.
38. Baron G, Tubach F, Ravaud P, ILogeart I, Dougados M. Validation of a short form of the Western Ontario and McMaster Universities' Osteoarthritis Index function subscale in hip and knee osteoarthritis. Arthritis Care Res 2007;57(4):633–8.
39. Feinstein SR. Clinimetrics. New Haven: Yale University; 1987.
40. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998 Feb;25(2): 198–9.